The Stack: 3 TB of permissively licensed source code

Year: 2022 | Citations: 389 | Authors: Denis Kocetkov, Raymond Li, Loubna Ben Allal, Jia Li, Chenghao Mou

---

Abstract

Large Language Models (LLMs) play an ever-increasing role in the field of Artificial Intelligence (AI)--not only for natural language processing but also for code understanding and generation. To stimulate open and responsible research on LLMs for code, we introduce The Stack, a 3.1 TB dataset consisting of permissively licensed source code in 30 programming languages. We describe how we collect the full dataset, construct a permissively licensed subset, present a data governance plan, discuss limitations, and show promising results on text2code benchmarks by training 350M-parameter decoders on different Python subsets. We find that (1) near-deduplicating the data significantly boosts performance across all experiments, and (2) it is possible to match previously reported HumanEval and MBPP performance using only permissively licensed data. We make the dataset available at https://hf.co/BigCode, provide a tool called"Am I in The Stack"(https://hf.co/spaces/bigcode/in-the-stack) for developers to search The Stack for copies of their code, and provide a process for code to be removed from the dataset by following the instructions at https://www.bigcode-project.org/docs/about/the-stack/.